MÉTODOS, EXPERIENCIAS Y HERRAMIENTAS PARA EL APRENDIZAJE EXPERIENCIAL DE LA CIENCIA DE DATOS

Emilio Serrano[1], Daniel Manrique, Martin Molina y Luis Baumela

Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
{emilioserra, dmanrique, mmolina, lbaumela}@fi.upm.es

***Resumen**. La Ciencia de Datos es una revolución que ya está cambiando la manera en la que nos ocupamos de negocios, sanidad, política, educación e innovación. Hay una gran variedad de cursos online, másteres, grados, y asignaturas que se enfocan a la enseñanza de este campo interdisciplinar, donde existe una demanda creciente de profesionales. En este proyecto se estudian extensiones de las metodologías de aprendizaje experiencial para Ciencia de Datos, se proponen diferentes modelos de enseñanza y aprendizaje para esta ciencia, y se exploran posibles plataformas software para facilitar el seguimiento de estos modelos. Los modelos han sido empleados en una asignatura de Deep Learning, dentro del contexto de un máster internacional en Ciencia de Datos.*

Palabras clave: Aprendizaje Activo, Aprendizaje Experiencial, Big Data, Elaboración material docente, Máster, Metodología Aprendizaje Basado en Problemas,
Metodología Aprendizaje Orientado a Proyectos, Metodología Trabajo en Equipo/Grupo.

1. Introducción

La Ciencia de datos o *Data Science* (DS) es un campo interdisciplinar que se encarga de la extracción del conocimiento de los datos. Esta disciplina es particularmente compleja ante el Big Data: grandes volúmenes de datos que dificultan su almacenamiento, procesamiento y análisis con tecnologías estándar de las Ciencias de la computación. La Ciencia de los datos es una revolución que ya está cambiando la forma de hacer negocios, la sanidad, la política, la educación y la innovación [1].

La existencia de distintos repositorios de datos sobre los que construir conocimiento ofrece un caldo de cultivo privilegiado para diseñar un curso de DS como una serie de experiencias en problemas del mundo real. Pocos campos permiten al estudiante ponerse en la piel de perfiles tan diversos e interesantes como DS: economistas, administradores de empresas, médicos, biólogos, administradores de sitios webs, etcétera. De la misma manera, pocas disciplinas pueden ofrecer recompensas tan atractivas para el aprendizaje experiencial (EL) como los tres millones de dólares que obtuvo el ganador del concurso para predecir los pacientes que eran admitidos en un hospital estadounidense en el siguiente año; o el millón de dólares con el que la compañía Netflix premió al mejor predictor de valoraciones de películas.

Este proyecto de innovación se plantea para desarrollar métodos, experiencias y herramientas para el aprendizaje experiencial de DS [2]; y más específicamente, de una de las disciplinas de las que se nutre la Ciencia de los datos: el Deep Learning (DL) [3].

2. Métodos

Gran parte del desarrollo de la teoría del EL en los últimos treinta años han ganado relevancia por las investigaciones de David A. Kolb [4] donde sintetiza los principios de su propuesta de aprendizaje. Muchas publicaciones, artículos, y estudios de investigación han explorado el poder explicativo y la utilidad de su teoría en varias disciplinas y profesiones. En el ámbito de este proyecto de investigación, el ciclo de Kolb

[1] Coordinador del PIE.

es revisado e instanciado para el campo específico de DS. Aunque hay recientes trabajos proponiendo el EL para campos relacionados como las tecnologías de la información [5], este es el primer proyecto que se centra en el campo de DS [6].

Algunos ejemplos prácticos en el uso de este marco de trabajo teórico para las cuatro principales fases del ciclo de Kolb son descritas a continuación. Sobre la presentación de una *experiencia concreta*, los problemas difíciles de resolver con el conocimiento actual de los estudiantes puede ser un buen punto de partida. Como el reconocimiento de caras o imágenes usando técnicas de ciencias de la computación sin inteligencia artificial. Esto puede motivar la necesidad de DS. En la *observación reflexiva* algunas propuestas [2] incluyen preguntas como *"¿Has notado que…?"*. Por ejemplo, *"¿Has notado que Facebook reconoce las caras en las fotos para que puedas etiquetarlas?".* En la *conceptualización abstracta*, se puede solicitar a los estudiantes modificaciones sobre los métodos ya estudiados a cierto nivel de abstracción. Por ejemplo, el uso de diagramas de flujo si la programación de las soluciones es demasiado exigente. Con respecto a la *experimentación activa*, se puede proponer la participación en competiciones académicas de aprendizaje automático como las publicadas regularmente en Kaggle [7].

3. Experiencias

En esta sección se describen los tres modelos propuestos para la enseñanza y el aprendizaje en DS, y más concretamente DL.

Redes de neuronas artificiales (ANN)

Las ANN son modelos computacionales inspirados en la neurociencia capaces de predecir una salida a partir de datos etiquetados (*aprendizaje supervisado*), así como de encontrar estructuras subyacentes y ocultas en datos no etiquetados (*aprendizaje no supervisado*) [8]. Tras su introducción, se dan consejos prácticos en la solución de problemas con ANN y se presentan dos entornos de trabajo de DS: Weka [8] y H20.ai. También se propone una experiencia a los estudiantes: entrenar y evaluar arquitecturas neuronales para un problema concreto, proporcionando un conjunto de datos que permita inferir un modelo inteligente de predicción, como los que se encuentran disponibles en el repositorio UCI.

Visión artificial (CV)

Uno de los aspectos fundamentales de este modelo es que se busca que los estudiantes comprueben que las lecciones aprendidas en ANN no permiten resultados aceptables para la visión artificial. En consecuencia, tras una introducción al campo, se propone a los estudiantes predecir el objeto que contiene una imagen. Además, esta tarea se propone en la modalidad de concurso para incluir un esquema de *gamificación*: el ranking de las mejores predicciones se muestra en la plataforma virtual de la asignatura. Gracias a esto, los estudiantes pueden reflexionar sobre los problemas que llevan a las ANNs a tener bajos porcentajes de predicciones correctas. Ningún estudiante superó el 57% de aciertos. Tras este concurso, se introducen las *redes de neuronas convolucionales* (*ConvNets*) y; en un enfoque práctico, se explica un entorno de trabajo para el diseño, entrenamiento, y uso de estas redes llamado Caffe. Después de introducir las ConvNets, se vuelve a plantear la misma experiencia en un nuevo concurso. Esta vez algunos estudiantes alcanzaron un 80% de predicciones correctas. Finalmente, se plantea una tercera experiencia en la que se reduce considerablemente los datos de entrenamiento, pero se permite utilizar ConvNets ya entrenadas para ajustarlas a los nuevos datos, i.e. *aprendizaje por trasferencia* en terminología de DS.

Procesamiento del Lenguaje Natural (NLP)

Para este modelo de enseñanza y aprendizaje, se invierte el orden de los modelos anteriores: primero se da una experiencia realista a los estudiantes y luego se ofrecen algunas soluciones alternativas. Más concretamente, se trata el problema de la *clasificación de texto*: se busca un modelo predictivo que sea capaz de catalogar un texto en una clase específica. En esta experiencia, los estudiantes forman grupos de hasta cinco personas, a modo de grupos de investigación, en un laboratorio equipado con un puesto de

trabajo por persona. Su tarea consiste en realizar una propuesta para una convocatoria competitiva de un proyecto de investigación en una gran compañía. El problema de investigación consiste en predecir la relevancia del título de un artículo respecto al cuerpo de este, basándose en un conjunto de datos con tres atributos: título, cuerpo, y clase (relevante/irrelevante). Además, para facilitar la exploración de diversas soluciones innovadoras, en lugar de pedir una implementación de código como en los modelos anteriores, se solicitan diagramas de flujo de trabajo para las propuestas, así como los entornos de trabajo que se van a usar y referencias bibliográficas. Posteriormente, los estudiantes deben revisar un extracto de lecturas seleccionadas (unas 5 páginas) con una introducción al campo del NLP, al problema de la clasificación de textos, y a plataformas de desarrollo de DL. Finalmente, se imparte una clase magistral donde se tratan diversas soluciones para el problema planteado. También se contextualizan estas soluciones con las distintas propuestas realizadas por los estudiantes.

Resultados

Al final del curso, se suministró un cuestionario a los 24 estudiantes presentes de un total de 29 matriculados en la asignatura DL. Para cada uno de los modelos, se plantearon tres afirmaciones: *A1* "El contenido del curso satisface mis necesidades de formación", *A2* "Lo que he aprendido será aplicable a mi trabajo", *A3* "La metodología aplicada, los recursos técnicos, y los materiales docentes eran apropiados". A los participantes se les pidió que calificaran su acuerdo o desacuerdo en una escala Likert de 5 puntos. Estas preguntas son similares a las realizadas por Alonso et al. [9]. La **¡Error! No se encuentra el origen de la referencia.** muestra la estadística descriptiva para la encuesta descrita [10], en donde se puede observar que, en todos los casos, la media de las respuestas dadas a las tres afirmaciones está por encima de la mitad de la escala (3).

Tabla 1. Evaluación de modelos de enseñanza y aprendizaje

	Media	Desviación típica
ANN		
A1	3,38	1,06
A2	3,58	0,78
A3	3,08	0,97
CV		
A1	4,50	0,59
A2	4,46	0,51
A3	4,00	0,88
NLP		
A1	4,13	0,68
A2	4,08	0,97
A3	4,21	0,93

4. Herramientas

En el contexto del proyecto se plantea el desarrollo, preferentemente reutilizando componentes o sistemas existentes, de una plataforma para el aprendizaje experiencial en ciencia de datos. Los repositorios de conjuntos de datos son muy usados en la docencia de DS, tales como el repositorio UCI o la lista de datasets de Kaggel [7].

Kaggel ofrece una inmensa base de datos con conjuntos de datos sobre los que utilizar DS, tutoriales básicos, foros de discusión para cada dataset, un listado de actividad reciente y *Kernels*. Los *Kernels* de Kaggle son un entorno de programación en la nube, reproducible y colaborativo, para el análisis de datos. Aunque Kaggle es una buena aproximación a la plataforma de experiencias de DS, no favorece la reflexión sobre los resultados y los paradigmas de aprendizaje tienden a verse como cajas negras de las que sólo importa la métrica de calidad conseguida. Algunos requisitos que este proyecto pretende cubrir

en su plataforma para aprendizaje experiencial en DS incluyen, entre otros: soporte a la decisión en la creación de experiencias en lugar de meros datasets, un sistema de reserva de recursos para permitir experiencias distinguidas entre distintos grupos de estudiantes, y un sistema de calificación privada para permitir a los Formadores evaluar la dificultad de las experiencias.

5. Conclusiones

En este proyecto se tratan métodos, experiencias y herramientas para potenciar el aprendizaje experiencial (EL) en la ciencia de datos (DS). Aunque existen trabajos recientes y relacionados en las tecnologías de la información, este proyecto parece ser el primero que trata EL en DS.

Observadas una serie de limitaciones en las técnicas pedagógicas más utilizadas en cursos de DS, el proyecto propone EL como forma de favorecer la reflexión y evitar la falta de pensamiento crítico. Para ello, se estudia y extiende el ciclo de Kolb a las particularidades de DS. Además, en el ámbito de un curso de Deep Learning, se han propuesto tres modelos de enseñanza y aprendizaje que permiten a los estudiantes obtener nuevo conocimiento iterando sobre experiencias realistas, reflexión sobre ellas, conceptualización del nuevo conocimiento y experimentación. Los tres modelos han recibido evaluaciones satisfactorias. Finalmente, se exploran algunos recursos populares en la enseñanza de DS y se exploran requisitos para una plataforma de experiencias que facilite los modelos de enseñanza y aprendizaje propuestos.

Tanto las encuestas anónimas realizadas como el material empleado en las experiencias propuestas en el proyecto están disponibles para el lector interesado solicitándolo a los autores. Igualmente, se espera extrapolar estos modelos a cursos con estudiantes de perfiles distintos y menos técnicos que los involucrados en este proyecto. Más concretamente, se ha planificado su uso en el Máster en Biología Computacional, ~~en~~ la Universidad Politécnica de Madrid.

REFERENCIAS

[1] V. Mayer-Schonberger y . K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.

[2] M. Jacobson y M. Ruddy, Open to Outcome: A Practical Guide for Facilitating & Teaching Experiential Reflection, Wood 'N' Barnes, 2004.

[3] I. Goodfellow, Y. Bengio y A. Courville, Deep Learning, MIT Press, 2016.

[4] D. A. Kolb. Experiential Learning: Experience as the Source of Learning and Development. Prentice Hall, 1 edition, Oct. 1983.

[5] S. Shiralkar, IT Through Experiential Learning: Learn, Deploy and Adopt IT through Gamification, Apress, 2016.

[6] E. Serrano, M. Molina, D. Manrique, L. Baumela, «Experiential learning in Data Science: from the dataset repository to the platform of experiences» Citizen-Centric Smart Cities Services (CCSCS'2017), at the 13th International Conference on Intelligent Environments (IE'17). To Appear, 21-22 Agosto, Seoul, Korea.

[7] Kaggle: Academic Machine Learning Competitions. https://inclass.kaggle.com/. Accessed: July of 2017.

[8] I. Witten, E. Frank y M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier Science, 2011.

[9] F. Alonso, G. López , J. M. Font y D. Manrique , «Learner satisfaction when applying an instructional model in e-learning: an experimental study» *Proceedings of the 2nd International Conference on Computer Supported Education - Volume 1: CSEDU,* pp. 141-146, 2010.

[10] E. Serrano, M. Molina, D. Manrique, L. Baumela y D. Zanardini, «Experiential learning in data science: student satisfaction for three models of teaching and learning» IV Congreso Internacional sobre Aprendizaje, Innovación y Competitividad (CINAIC 2017). To Appear, 4-6 Octubre, Zaragoza, Spain.